УДК 519.6:004.93

А.А. Олейник, С.А. Субботин
Запорожский национальный технический университет, Украина
subbotin@zntu.edu.ua

Мультиагентный метод с непрямой связью между агентами для выделения информативных признаков

Решена задача отбора информативных признаков. Предложены мультиагентные методы отбора информативных признаков с непрямой связью между агентами на основе представления пунктов назначения признаками и в виде информативности. Проведены эксперименты по решению задачи отбора информативных признаков для моделирования коэффициента упрочнения деталей газотурбинных авиадвигателей после алмазного выглаживания.

Введение

Главной целью отбора информативных признаков является выделение минимального подмножества признаков из проблемной области, которое обеспечивало бы требуемую точность описания экземпляров относительно набора первоначально заданных признаков [1]. При решении прикладных задач отбор информативных признаков необходим для исключения нерелевантных, малоинформативных признаков, а также признаков, являющихся следствием шумов. За счёт удаления таких признаков работа методов обработки данных, которые используются при моделировании сложных объектов и систем, может выполняться значительно быстрее.

Постановка задачи

Пусть задано множество признаков $X = \{x_1, x_2, ..., x_n\}$ размерности n. Тогда задача отбора информативных признаков [2] заключается в поиске такого набора признаков X^* из всего возможного множества 2^n наборов-кандидатов, критерий оптимальности для которого находился бы в допустимых пределах:

$$\varepsilon(X^*) \le \varepsilon_n ,$$

где $\varepsilon(X^*)$ – значение критерия оптимальности для набора признаков X^*; ε_n – пороговое значение критерия оптимальности.

Обычно методы отбора информативных признаков используют либо технику эвристического поиска [1], [2], либо технику случайного поиска [1], [2] для снижения вычислительной сложности. Однако при таком подходе часто уменьшается уровень оптимальности отобранного набора информативных признаков. В связи с этим в данной работе предлагается применять мультиагентный метод с непрямой связью между агентами [3], [4], поскольку он позволяет решать различные задачи дискретной оптимизации, к которым может быть отнесена и задача отбора информативных признаков. При этом оптимизационный процесс характеризуется меньшей вычислительной сложностью, чем у методов полного или сокращённого перебора, а также мультиагентный метод с непрямой связью между агентами обеспечивает требуемую точность результирующего решения [3-9].

Применение метода муравьиных колоний к решению задачи отбора информативных признаков

Метод муравьиных колоний является эвристическим, итеративным методом случайного поиска, основанным на моделировании поведения агентов (муравьёв) в процессе решения ими оптимизационных задач [3], [4]. Данный метод основан на мультиагентном подходе с непрямой связью между агентами, которая достигается за счёт моделирования меры приоритетности граней в графе поиска. Метод широко применяется для решения различных задач дискретной оптимизации, например, задачи коммивояжёра, задачи календарного планирования и т.п. [5-8]. Однако в общем случае метод муравьиных колоний может быть применен к любой комбинаторной задаче, которая согласуется со следующими требованиями:

– соответствующее представление задачи – задача должна быть описана в виде графа с набором узлов и граней между узлами;

– эвристическая пригодность граней – возможность применения эвристической меры адекватности путей от одного узла к каждому соседнему узлу в графе;

– составление альтернативных решений, посредством чего можно рационально определять допустимые решения;

– правило обновления феромонов – правило, которое определяет вероятность перемещения агента из одного узла графа к другому.

Задача отбора информативных признаков может быть представлена в виде, соответствующем методу муравьиных колоний. Модификации метода муравьиных колоний для решения задачи отбора информативных признаков заключаются в том, что задача представляется в виде графа, узлы которого характеризуют признаки, а грани между ними – выбор следующего признака.

Для решения задачи отбора информативных признаков предлагается использовать два метода, основанные на мультиагентном подходе с непрямой связью между агентами: мультиагентный метод на основе представления пунктов назначения признаками и метод на основе представления пунктов назначения в виде информативности признаков.

Мультиагентный метод отбора информативных признаков с непрямой связью между агентами на основе представления пунктов назначения признаками

Основная идея данного метода в аспекте решения задачи отбора признаков заключается в следующем: предполагается, что агент должен совершить путь по заданному количеству пунктов назначения n, при этом каждому пункту ставится в соответствие признак x_i, $i = \overline{1,n}$; тогда путь, пройденный агентом, составляет набор признаков H, который предлагается использовать для построения модели. При этом данные хранятся не в битовом виде, а в виде номеров отобранных признаков $j = \overline{1,n}$. При получении итогового результата это позволяет получать оценку степени влияния каждого признака.

Шаг 1. Задать параметры метода: *initPh* – начальное количество феромона; ρ – коэффициент количества феромона, которое агент оставляет на пути, где $(1 - \rho)$ показывает коэффициент испарения феромона на пути после его завершения; Q – константу, относящуюся к количеству феромона, которое было оставлено на пути; *inCF* – количество признаков в исходном наборе; *outCF* – количество признаков, которое следует оставить в сокращённом наборе.

Шаг 2. Инициализация. Создание популяции агентов. После создания популяция агентов поровну распределяется по пунктам назначения.

Шаг 3. Движение агентов. Если агент еще не закончил путь, то есть не посетил пункты в количестве, равном *outCf*, для определения следующего пункта в пути агента используется вероятность:

$$P_k = \frac{\tau_k(t)}{\sum_{i=1}^{n^j} \tau_i^j(t) + \tau_k(t)},$$

где P_k – вероятность того, что j-й агент, который ещё не посетил n^j пунктов, для продолжения пути выберет пункт k; $\tau_i(t)$ – количество феромона в i-ом пункте в момент времени t.

Агент перемещается только в пункты, которые еще не были посещены (как указано списком табу *tList*). Поэтому вероятность P_k рассчитывается только для пунктов, которые ещё не были посещены агентом.

Шаг 3 повторяется до тех пор, пока каждый агент не закончит свой путь. Путь считается пройденным до конца, когда агент посетит пункты в количестве, равном *outCF*.

Шаг 4. На основе пунктов, посещённых агентом, строится модель $F(X^{\text{отобр.}})$ и для неё определяется ошибка ε_j.

Шаг 4.1. Последовательность номеров узлов, посещённых агентом, переводится в битовую строку H^j по формуле:

$$a_i = \begin{cases} 0, \text{ если } i \notin L^j; \\ 1, \text{ если } i \in L^j, \end{cases}$$

где a_i – i-й бит в битовой строке; i – номер узла; L^j – путь j-го агента.

Шаг 4.2. На основе полученной битовой строки H^j и экспериментальных данных строится модель (например, на основе регрессии или нейросети).

Шаг 4.3. Рассчитывается ошибка ε_j для полученной модели:

$$\varepsilon_j = \frac{1}{2}\sum_{i=1}^{N}(y_i - y_{i\ расч})^2,$$

где y_i – исходные значения выходной переменной; $y_{i\ расч}$ – расчётное значение выходной переменной по построенной модели; N – количество экспериментов.

Шаг 5. Определяется количество феромона, которое было оставлено в каждом пункте пути для j-го агента:

$$\Delta \phi^j(t) = \frac{Q}{e_j},$$

где $\Delta\tau^j(t)$ – количество феромонов, которое надо добавить каждому пункту, входящему в путь j-го агента в момент времени t; Q – параметр, понижающий степень влияния ошибки модели ε_j, построенной на основании признаков, входящих в путь j-го агента.

Затем $\Delta\tau^j(t)$ используется для расчёта того, насколько требуется увеличить количество феромона для каждого пункта, в котором побывал агент:

$$\tau_i(t) = \tau_i(t-1) + \rho \cdot \sum_{j=1}^{N} \Delta\tau_i^j(t),$$

где N – количество агентов, выбравших i-й признак.

Данная формула применяется ко всему пути, при этом каждый пункт помечается феромоном пропорционально ошибке полученной модели. Поэтому следует дождаться, пока агент закончит путешествие, и только потом обновить уровни феромона, в противном случае истинная длина пути останется неизвестной.

В начале пути у каждого пункта есть шанс быть выбранным. Чтобы постепенно удалить пункты, которые входят в худшие комбинации, ко всем путям применяется процедура испарения феромона:

$$\tau_i(t) = \tau_i(t) \cdot (1 - \rho) .$$

Шаг 6. Проверка на достижение оптимального результата. Результат считается оптимальным, если:

– достигнуто максимальное количество итераций: $t = t_{\max}$;

– полученный результат находится в допустимом пределе: $\varepsilon_{cur} = \varepsilon_{threshold}$;

– на протяжении cnt_t^{repeat} итераций не было отмечено изменений в выборе наилучшего решения.

Если проверка дала положительный результат, то происходит окончание работы метода (переход к шагу 7), в противном случае – переход к шагу 6.

Шаг 7. Повторный запуск. После того, как путь агента завершен, грани обновлены в соответствии с длиной пути и произошло испарение феромона на всех гранях, метод выполняется повторно. Список табу очищается, и длина пути обнуляется. Агентам разрешается перемещаться по сети. Переход к шагу 3.

Шаг 8. Останов. Определяется лучший путь, который является решением. Лучший путь выбирается путём сравнения ошибок моделей, которые рассчитываются по принципу, описанному на шаге 4.

Шаг 9. После останова и выбора наилучшего пути строится финальная модель, что позволяет обеспечить требуемую точность при меньшем количестве признаков, чем в исходном наборе.

Таким образом, предложенный метод на основе представления пунктов назначения признаками позволяет решать задачу отбора информативных признаков за счёт приведения задачи отбора информативных признаков в вид, соответствующий концепции работы мультиагентного подхода с непрямой связью между агентами. При этом основной особенностью предложенного метода является то, что пункт назначения в графе поиска представляет собой признак, который предлагается принять как информативный. Поскольку в процессе работы мера приоритетности каждого узла изменяется, то за счёт такого подхода можно выделить не только наиболее информативные признаки, а также и выяснить относительную важность каждого из них.

Мультиагентный метод отбора информативных признаков на основе представления пунктов назначения в виде информативности признаков

Особенностью данного метода является то, что пунктам назначения $(\overline{1,n})$, где n – количество входных признаков *inCF*, ставится в соответствие двоичный набор $B = \{b_i \mid b_i = <0;1>, i = \overline{1,n}\}$, формируемый случайно, и при этом количество единичных элементов должно быть равным количеству признаков *outCF*, которые необходимо оставить. Каждый агент должен совершить путь по всем пунктам назначения, после чего по полученному пути строится модель в соответствии с двоичным набором B.

Основные шаги данной модификации приведены ниже.

Шаг 1. Задать параметры метода: α – коэффициент, определяющий относительную значимость пути (количество феромона на пути); *initPh* – начальное количество феромона; ρ – коэффициент количества феромона, которое агент оставляет на пути, где (1– ρ) показывает коэффициент испарения феромона на пути после его заверше-

ния; Q – константу, относящуюся к количеству феромона, которое было оставлено на пути; $inCF$ – количество признаков в исходном наборе; $outCF$ – количество признаков, которое следует оставить в сокращённом наборе, $maxTime$ – максимальное время моделирования. Установить время моделирования $curTime$ в 0.

Шаг 2. Инициализация. Создание двоичного набора B и популяции агентов. Двоичный набор B следует создавать случайным образом. Предлагается получить $outCF$ целых случайных чисел $rNum_j$ ($j = \overline{1,outCF}$) в интервале [1; $inCF$] и каждому b_{rNum_j} присвоить 1, остальные элементы b_k принять равными 0. Количество агентов в популяции $agentCnt$ предлагается выбирать не меньшим количества признаков $inCF$. После создания популяция агентов поровну распределяется по пунктам назначения.

Шаг 3. Изменение количества времени моделирования: $curTime = curTime + 1$.

Шаг 4. Движение агентов. Каждый агент, пока не закончит весь путь, выбирает следующий пункт назначения. При этом при выборе следующего пункта учитывается количество феромона, уже оставленного в этом пункте, и количество феромона, оставленного в других пунктах, ещё не посещённых агентом. Добавление пункта k в путь j-го агента происходит с вероятностью:

$$P_k^j(t) = \frac{\tau_k(t)}{\sum_{i=1}^{n^j} \tau_i^j(t) + \tau^\alpha{}_k(t)},$$

где $\tau_i^j(t)$ – количество феромона в i-м пункте j-го агента, который ещё не посетил n^j пунктов.

Шаг 5. Если каждый агент посетил все пункты, то происходит обновление путей (шаг 6) и перезапуск агентов (шаг 7). В противном случае – переход к шагу 3.

Шаг 6. Обновление путей. Данный этап необходим для того, чтобы понизить вероятность в дальнейшем выбора таких пунктов назначения, включение которых даёт повышение ошибки построенной модели относительно других моделей, построенных без этого пункта.

Шаг 6.1. Для каждого агента строится модель на основе пунктов, посещённых им. При этом происходит декодирование из последовательности номеров пунктов в двоичный код для определения тех признаков, которые следует оставить для построения модели. Декодирование происходит исходя из двоичного набора B. Каждому пункту x^j_i ($i = \overline{1,\text{outCF}}$) j-го агента ставится в соответствие число a_i по следующему принципу:

$$a_i = \begin{cases} 1, \text{ если } b_{x_i^j} = 1; \\ 0, \text{если } b_{x_i^j} = 0. \end{cases}$$

Шаг 6.2. На основе полученной битовой строки $H^j = \{a_i, i = \overline{1,outCF}\}$ и экспериментальных данных строится модель (например, на основе регрессии или нейросети).

Шаг 6.3. Рассчитывается ошибка ε_j:

$$\varepsilon_j = \frac{1}{2}\sum_{i=1}^{N}(y_i - y_{i\,расч})^2,$$

где y_i – исходные значения выходной переменной; $y_{i\,расч}$ – расчётное значение выходной переменной по построенной модели; N – количество экспериментов.

Шаг 6.4. В соответствии с рассчитанной ошибкой количество феромонов каждого пункта, который посетил j-й агент, увеличивается на $\Delta\tau(j)$:

$$\phi_i^j(t) = \Delta\phi^j(t) + (\phi_i^j(t)\cdot\rho),$$

где $\tau^j_i(t)$ – количество феромона в i-м пункте j-го агента;

При этом $\Delta\tau^j(t)$ определяется исходя из рассуждений, что больше феромона должно добавляться тем пунктам, использование которых даёт меньшую ошибку, при этом следует влияние значения самой ошибки на количество добавляемого феромона сделать не слишком существенным, поскольку может возникнуть зацикливание на одном пути:

$$\Delta\tau^j(t) = \frac{Q}{\varepsilon_j},$$

где Q – параметр, понижающий степень влияния ошибки модели ε_j, построенной на основании признаков, входящих в путь j-го агента.

Шаг 6.5. В начале пути у каждого пункта есть шанс быть выбранным. Чтобы постепенно удалить пункты, которые входят в худшие комбинации, ко всем путям применяется процедура испарения феромона:

$$\tau_i(t) = \tau_i(t) \cdot (1 - \rho).$$

Шаг 7. Перезапуск агентов. На данном шаге производятся действия, аналогичные шагу 2. Однако в данном случае уже имеющаяся популяция распределяется по начальным пунктам, обнуляется список табу *tList*. Также на этом шаге выбирается лучший из путей, которые были получены при данном запуске агентов, и сохраняется в *bestPath*.

Шаг 8. Проверка на останов. Осуществляется либо на основе времени моделирования *curTime* (останов происходит при достижении предела *maxTime*), либо на основании ошибки лучшей из полученных моделей $(\varepsilon_{bestPath} < \varepsilon)$. Если проверка на останов дала положительный результат – переход к шагу 9, в противном случае – переход к шагу 3.

Шаг 9. Останов. Определяется лучший путь, который является решением. Лучший путь выбирается на основе ошибок моделей, которые рассчитываются по принципу, описанному на шаге 4.

Шаг 10. После останова и выбора наилучшего пути строится финальная модель, что позволяет обеспечить требуемую точность при меньшем количестве признаков, чем в исходном наборе.

Таким образом, предложенный мультиагентный метод с непрямой связью между агентами на основе представления пунктов назначения в виде информативности отличается от предыдущего метода тем, что в данном методе узлами являются информативности, за счёт чего обеспечивается более детальное изучение пространства поиска, поскольку при таком подходе обеспечивается рассмотрение большего количества различных комбинаций признаков.

Эксперименты и результаты

Предложенные методы были программно реализованы на языке программирования среды Matlab. С помощью разработанного программного обеспечения решалась задача прогнозирования коэффициента упрочнения деталей авиадвигателей после алмазного выглаживания [10], [11]. Также для сравнения работы предложенных методов с другими методами данная задача решалась с помощью канонического метода эволюционного поиска [12].

В качестве факторов, наиболее полно отражающих процесс алмазного выглаживания деталей, в [10], [11] предлагается использовать: x_1 – твердость материала, НВ, МПа; x_2 – среднее контактное давление в зоне деформирования, q_{max}, МПа; x_3 – полуось эллипса касания в зоне упругого контакта, a, мм; x_4 – подача при выглаживании, s, мм/об, x_5 – предел прочности, $\sigma_В$, МПа; x_6 – предел текучести материала, $\sigma_{0.2}$, МПа; x_7 – показатель деформационного упрочнения, n; x_8 – теоретический коэффи-

циент концентрации напряжений от следов обработки, $\alpha_{\sigma}^{\text{техн}}$; x_9 – параметр исходной шероховатости детали, R_{a1}, мкм; x_{10} – сила выглаживания, P_y, H; x_{11} – радиус сферы алмазного инструмента, $R_{сф}$, мм; x_{12} – параметр шероховатости после выглаживания детали, R_{a2}, мкм; x_{13} – теоретический коэффициент концентрации напряжений натурной детали (образца), α_{σ}; x_{14} – диаметр образца в опасном сечении, d, мм; x_{15} – радиус скругления галтели или надреза, r, мм; x_{16} – относительный градиент первого главного напряжения, $\overline{G}$, мм$^{-1}$.

В [11] были построены статистические модели коэффициента упрочнения, полученные на основе теории подобия. Ошибка таких моделей составила порядка 10% при расчёте коэффициента упрочнения [11].

С помощью разработанного программного обеспечения было получено несколько комбинаций наиболее информативных признаков, среди которых были отобраны две лучшие комбинации: (x_3, x_5, x_7, x_8, x_{11}, x_{14}, x_{16}) и (x_1, x_7, x_8, x_{10}, x_{11}, x_{12}, x_{16}), полученные с помощью мультиагентного метода на основе представления пунктов назначения узлами и на основе представления пунктов назначения в виде информативности соответственно. Среднее значение ошибки моделей, построенных на основе отобранных признаков, составило 5,18% и 5,05% соответственно, что значительно лучше по сравнению с ошибкой полученных ранее моделей [11].

Сравнение работы предложенных мультиагентных методов с непрямой связью между агентами для отбора информативных признаков с работой канонического метода эволюционного поиска представлено на рис. 1.

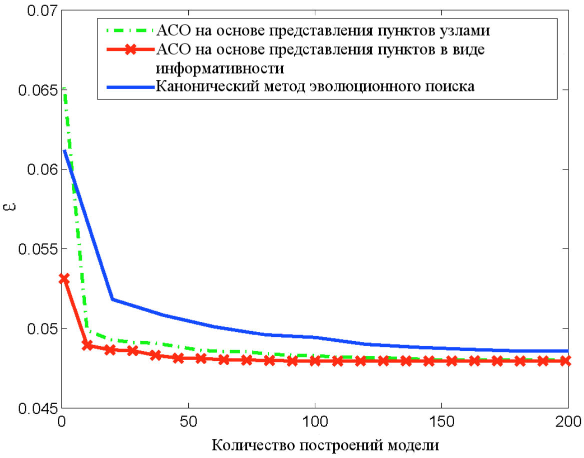


Рисунок 1 – Сходимость к оптимальному решению как функция ошибки прогнозирования модели ε от количества построений модели (количество запусков: 20, количество информативных признаков: 9)

Как можно видеть из рисунка, предложенные методы, основанные на мультиагентном подходе, сходятся к оптимальному решению значительно быстрее, чем канонический метод эволюционного поиска, что свидетельствует об эффективности предложенных мультиагентных методов с непрямой связью между агентами для отбора информативных признаков.

Заключение

В работе решена актуальная задача отбора информативных признаков на основе мультиагентного подхода с непрямой связью между агентами.

Научная новизна работы заключается в том, что впервые предложен мультиагентный подход с непрямой связью между агентами для отбора информативных признаков,

который за счёт представления пунктов назначения либо признаками, либо информативностями, а также за счёт назначения меры приоритетности узлам, а не граням, позволяет обеспечить невысокую вычислительную сложность оптимизационного процесса, достигая при этом детального исследования пространства поиска, и как следствие, достигая необходимой точности прогнозирования моделей, полученных на основе отобранных наборов информативных признаков.

Практическая ценность результатов работы состоит в том, что разработано программное обеспечение, реализующее предложенные методы отбора информативных признаков на основе мультиагентного подхода с непрямой связью между агентами, на основе которого решена задача моделирования коэффициента упрочнения деталей газотурбинных авиадвигателей после алмазного выглаживания.

Литература

1. Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей: монография / В.И. Дубровин, С.А. Субботин, А.В. Богуслаев, В.К. Яценко. – Запорожье : ОАО «Мотор-Сич», 2003. – 279 с.
2. John G.H. Irrelevant Features and the Subset Selection Problem / G.H. John, R. Kohavi, K. Pfleger // Proceedings of the Eleventh International Conference on Machine Learning. – San Francisco : Morgan Kaufmann Publishers, 1994. – P. 121-129.
3. Dorigo M. Optimization, Learning and Natural Algorithms / Dorigo M. – Milano : Politecnico di Milano, 1992. – 140 p.
4. Субботін С.О. Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей: монографія / С.О. Субботін, А.О. Олійник, О.О. Олійник ; [під заг. ред. С.О. Субботіна]. – Запоріжжя : ЗНТУ, 2009. – 375 с.
5. Dorigo M. The Ant System: Optimization by a colony of cooperating agents / M. Dorigo, V. Maniezzo, A. Colorni // IEEE Transactions on Systems, Man, and Cybernetics. – 1996. – Part B, № 26 (1). – P. 29-41.
6. Gambardella L.M. Solving symmetric and asymmetric TSPs by ant colonies / L.M. Gambardella, M. Dorigo // Proceedings of the 1996 IEEE International Conference on Evolutionary Computation (ICEC'96). – NJ : IEEE Press, 1996. – P. 622-627.
7. Dorigo M. Ant Colony System: A cooperative learning approach to the traveling salesman problem / M. Dorigo, L.M. Gambardella // IEEE Transactions on Evolutionary Computation. – 1997. – № 1 (1). – P. 53-66.
8. Ant system for job-shop scheduling / A. Colorni, M. Dorigo, V. Maniezzo, M. Trubian // Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL). – 1994. – № 34. – P. 39-53.
9. Stützle T. The MAX–MIN Ant System and local search for the traveling salesman problem / T. Stützle, H.H. Hoos // Proceedings of the 1997 IEEE International Conference on Evolutionary Computation (ICEC'97) / Eds. : T. Back, Z. Michalewicz, and X. Yao. – NJ : IEEE Press, 1997. – P. 309-314.
10. Яценко В.К. Повышение несущей способности деталей машин алмазным выглаживанием / В.К. Яценко, Г.З. Зайцев, В.Ф. Притченко. – М. : Машиностроение, 1985. – 232 с.
11. Богуслаев В. А. Технологическое обеспечение и прогнозирование несущей способности деталей ГТД / В.А. Богуслаев, В.К. Яценко, В.Ф. Притченко. – К. : Манускрипт, 1993. – 333 с.
12. Siedlecki W.A note on genetic algorithms for large-scale feature selection / W. Siedlecki, J. Sklansy // Pattern Recognition Letters. – 1989. – Vol. 10. – P. 335-347.

О.О. Олійник, С.О. Субботін

Мультиагентний метод з непрямим зв'язком між агентами для відбору інформативних ознак

Вирішено задачу відбору інформативних ознак. Запропоновано мультиагентні методи відбору інформативних ознак з непрямим зв'язком між агентами на основі представлення пунктів призначення ознаками та у вигляді інформативності. Проведено експерименти з вирішення задачі відбору інформативних ознак для моделювання коефіцієнту зміцнення деталей газотурбінних авіадвигунів після алмазного вигладжування.

A.A. Oleynik, S.A. Subbotin

Multiagent Method with Indirect Connection between Agents for Feature Selection

The feature selection problem is solved. Multiagent methods for feature selection with indirect connection between agents are proposed. The experiments to solving of problem of model synthesis of airengine details hardening coefficient are carried out.

Статья поступила в редакцию 03.06.2009.